基于卷积神经网络的视频图像超分辨率重建方法 *

刘　村，李元祥，周拥军，骆建华

(上海交通大学 航空航天学院, 上海 200240)

摘　要：为了进一步增强视频图像超分辨率重建的效果，研究利用卷积神经网络的特性进行视频图像的空间分辨率重建，提出了一种基于卷积神经网络的视频图像重建模型。采取预训练的策略用于重建模型参数的初始化，同时在多帧视频图像的空间和时间维度上进行训练，提取描述主要运动信息的特征进行学习，充分利用视频帧间图像的信息互补进行中间帧的重建。针对帧间图像的运动模糊，采用自适应运动补偿加以处理，对通道进行优化输出得到高分辨率的重建图像。实验表明，重建视频图像在平均客观评价指标上均有较大提升（PSNR +0.4 dB / SSIM +0.02），并且有效减少了图像在主观视觉效果上的边缘模糊现象。与其他传统算法相比，在图像评价的客观指标和主观视觉效果上均有明显的提升，为视频图像的超分辨率重建提供了一种基于卷积神经网络的新颖架构，也为进一步探索基于深度学习的视频图像超分辨率重建方法提供了思路。

关键词：视频；超分辨率重建；卷积神经网络；深度学习

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2017.10.1020

Video image super-resolution reconstruction method based on convolutional neural network

Liu Cun, Li Yuanxiang, Zhou Yongjun, Luo Jianhua

(*School of aeronautics & astronautics*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: In order to further improve the performance of video image super-resolution reconstruction and study the reconstruction of spatial resolution of video images by using the characteristics of convolution neural network, this paper proposed a video image reconstruction model based on convolution neural network. The model adopted the pre-training strategy to initialize the parameters. And it carried out the training processing both on the spatial and temporal dimensions of the multi-frame video images at the same time. It extracted the characteristics of the main motion information, learn and make full use of the information inter the frames for improved performance. And it used the adaptive motion compensation algorithm to optimize the output of the channel to obtain the reconstructed center frame image with high resolution. The experimental results show that the average of objective evaluation indexes for video image reconstruction has improved with a rather clear margin (PSNR + 0.4 dB / SSIM + 0.02) , and the edge of the fuzzy phenomenon in video reconstruction image for the subjective visual effect has been effectively reduced. Compared with other traditional algorithms, the evaluation of the objective indexes and subjective visual effect of the reconstructed image are both obviously improved. Provide a novel architecture based on convolution neural network for video image super-resolution, which provides an exploration for the further study of video image super-resolution reconstruction based on the deep learning method.

Key Words: video; super-resolution reconstruction; convolutional neural network; deep learning

0　引言

图像超分辨率重建是从低分辨率图像或视频序列中获得对应高分辨率图像的过程，在医学、航空和电子监控等诸多领域均有广泛应用[1]。随着新一代超高清视频（3840×2048）的日益普及，大多数视频内容在获取、传输和保存中面临许多问题，因此，需要视频重建算法从全高清（1920×1080）或更低分辨率的视频中生成超高清内容。

目前，图像超分辨率重建方法可以分为两类——基于模型的重建方法和基于学习的重建方法。基于模型的重建方法将低分辨率图像建模为具有随机噪声的高分辨率图像的二次采样图像，在从低分辨率图像恢复高分辨率图像的过程中通过引入正则化项，对模型进行约束[2]。在贝叶斯框架中，引入决定图像平滑度的先验知识，以获得质量更高的重建图像，例如，Babacan

基金项目：国家自然科学基金资助项目（11672183）；上海市军民融合专项资助项目（2016GFZ-GB02-342）

作者简介：刘村（1991-），男，山东淄博人，硕士研究生，主要研究方向为图像处理、深度学习（liucun_sjtu@sjtu.edu.cn）；李元祥（1967-），男，副教授，主要研究方向为遥感图像解译、图像识别、图像重构与评估；骆建华（1958-），男，教授，主要研究方向为超分辨率重建、医学图像处理．

等人[3]利用贝叶斯框架从多帧旋转和平移的低分辨率图像中重建得到高分辨率图像；Belekos 等人[2]以及 Liu 等人[4]也使用贝叶斯框架推导出一种能够处理复杂物体运动和实际场景视频序列的算法。

基于学习的重建算法的核心思想是学习高分辨率图像和对应低分辨率图像对之间的映射关系，其中学习字典由高分辨率图像块和低分辨率图像块联合训练所得，每个低分辨率图像块可以表示为来自对应低分辨率字典的原子的稀疏线性组合，字典中通过训练得到的系数表示权重[5]。字典及其权重系数可以用标准稀疏编码技术（如 K-SVD[6]）得到，然后可以通过找到对应的低分辨率图像块的稀疏系数，并将其应用于高分辨率字典中来重建对应高分辨率图像块。基于字典学习的重建方法中，自然图像块可以被稀疏地表示为学习字典块或原子的线性组合。Yang 等人[5]首先提出使用两个耦合字典来学习低分辨率图像和高分辨图像之间的非线性映射。Song 等人[7]提出一种视频超分辨率重建的字典学习方法——字典实时学习，其假设高分辨率图像中存在稀疏的关键帧是可用的。基于学习的图像重建方法通常要学习图像块的表示，可以通过图像块重建对应的高分辨率图像[8]。为避免重建图像中明显的边缘效应，算法中通常使用重叠的图像块进行运算，导致相当大的计算开销，影响算法的效率。Timofte 等人[9]提出使用几个较小的完整字典替换单个大型不完整字典，以降低算法中的稀疏编码步骤的计算开销，在保持重建精度的同时，能够进一步提高算法的速度。Schulter 等人[10]提出训练随机森林模型，而不是直接训练低分辨率图像块到高分辨率图像块的耦合字典。Glasner 等人[11]没有从样本图像中学习字典，而是用不同的缩放因子创建了一组低分辨率图像，然后将低分辨率图像中的图像块与自身的缩放版本进行匹配，使用其高分辨率图像块重建对应的高分辨率图像。

传统的单幅图像超分辨率重建主要着力于对图像的降质过程进行分析，从而在由低分辨率到高分辨率的重建过程中获取更多的建模信息，或者充分利用图像的自相似性，以及相关的变换域分析方法，在超分辨率重建的过程中恢复更多的高频信息。而对于视频图像的超分辨率重建问题，由于视频图像帧间的信息丢失、空间移位和翻转，或者帧间信息的冗余，在重建过程中需要着力解决帧间信息的匹配、多帧图像重建的预处理、方位估计和运动估计，以及运动模糊等问题，对降质模型的参数估计和重建也更加困难。

基于贝叶斯框架的视频重建方法采用更复杂的光流算法或分层块匹配方法来估计运动场，以便能够处理具有更复杂运动方式的真实场景视频[12]。Ma 等人[13]扩展了基于贝叶斯框架的工作，用于处理具有运动模糊的视频，并且引入时间相对度的概念，去除了严重模糊的像素，从而得到更优的实验效果。Takeda 等人[14]提出了常规运动估计和图像重建方案的另一种思路，直接利用 3-D 迭代转向内核回归，而不是简单的运动估计，视频能够被重叠的 3D 立方块（时间和空间维度）分割和处理，然后通过用 3D 泰勒级数逼近立方块中的像素来重建高分辨率图像。

随着深度学习以及卷积神经网络在图像识别领域的巨大成功，基于深层神经网络的新一代图像重建算法开始表现出卓越的性能[15]。深层神经网络能够学习和处理诸如 ImageNet 等大型训练数据库，可以通过使用 GPU 加速计算的并行化来有效地实现卷积神经网络的训练。此外，网络模型一旦训练完成，图像的超分辨率重建过程即为纯粹的前馈过程，这也使得基于卷积神经网络的重建算法在时间性能上表现更优[16]。Dong 等人[17]指出，基于字典学习的重建算法的每个步骤均可以被重新定义为深层神经网络中的一层，其中，用具有 n 个原子的字典表示尺寸为 $f \times f$ 的图像块，即在输入图像上应用具有核大小为 $f \times f$ 的 n 个滤波器，从而可以被实现为卷积神经网络中的一个卷积层。因此，其提出了一个卷积神经网络模型，通过使用具有两个隐含层和一个输出层的卷积神经网络结构直接学习从低分辨率图像到高分辨率图像的非线性映射。Wang 等人[18]引入了一种基于图像块的方法，其中卷积自动编码器用于 33×33 像素的图像块，减去均值的标准化低/高分辨率图像块用于训练重建模型，训练图像块根据其相似性进行聚类，并且对于每个类群，自相似的图像块和每一个子模型进行微调，充分利用图像的二维数据结构，同时训练数据增加了平移、旋转及不同的缩放因子等因素，以使模型能够更有效地学习视觉意义上的特征。Cui 等人[19]提出了一种将低分辨率图像的分辨率逐渐提高到所需分辨率的算法，它由层叠的协同局部自动编码器（CLA）组成，在级联的每层中执行非局部自相似性搜索（NLSS）以重建图像的高频细节和纹理，所得到的图像再由自动编码器处理，以消除由 NLSS 步骤引入的结构失真和错误。该算法使用 7×7 像素的重叠块，导致计算中的开销非常大。Cheng 等人[20]使用完全连接的网络层引入了基于图像块的视频重建算法，该网络包括隐含层和输出层，使用 5 帧连续的低分辨率图像帧重建一帧中心高分辨率图像框架，视频经过图像块处理，其中网络输入为 5×5×5 的图像块，并从高分辨率图像输出重建的图像块，通过使用参考帧和相邻帧匹配来找到对应的图像块。Liao 等人[21]提出了一种类似的方法，涉及多帧运动补偿并使用卷积神经网络组合帧，首先利用不同参数设置的两种运动补偿算法来计算重建图像的初始结果，以处理运动补偿误差；然后将所有初始版本的重建图像均使用卷积神经网络进行组合，但算法中每一帧需要计算多个运动补偿量，导致在计算开销上非常大。

综上分析，大多数视频重建算法依赖于低分辨率视频帧之间的精确运动估计，很多算法需要同时估计密集运动场和计算运动向量，不仅造成巨大的运算开销，而且需要在进行超分辨率重建之前增加相当复杂的处理步骤。

1　基于卷积神经网络的视频重建方法

本文提出一种新颖的基于卷积神经网络的视频重建模型，使用多个低分辨率视频帧作为输入，在输入前引入自适应运动补偿机制来处理视频连续帧中的快速移动目标以及由此造成的

运动模糊现象，同时对输出层进行通道优化，重建得到一幅高分辨率的输出帧图像。本文还针对不同的卷积网络模型结构作了比较，采用图像预训练策略，利用得到的权重系数初始化视频重建模型中的训练参数，在重建精度和速度方面均提高了模型的整体性能。

1.1　视频重建模型

基于学习的重建方法指出，在重建过程中使用相邻帧对于视频重建是有帮助的[22]。在视频重建的过程中，对帧间的运动进行建模和估计，可以通过帧间的子像素运动获得附加信息。如果训练过程中包含多个视频帧，则通过基于学习的方法也可以获得由帧间差异所传达的附加信息。

视频重建模型架构（图 1）主要包括一个运动补偿模块和三个卷积层（L_1、L_2 及 L_3）。本文将相邻帧（前一帧 $(t-1)$，当前帧 (t) 和后一帧 $(t+1)$）组合作为输入帧的结构，包含在整个重建过程中。为了使用多个前向和后向帧，模型架构可以扩展更多的分支。单帧输入的框架尺寸为 $1\times M\times N$，其中 M 和 N 分别为输入图像帧的宽度和高度。对于所提出模型的输入架构，在应用第一个卷积层 L_1 之前，三帧输入图像沿着第一维进行连接，那么 L_1 新输入的数据结构即为 $3\times3\times M\times N$ 维。类似的方式，模型也可以在卷积层 L_1 之后进行帧的连接组合，然后用做 L_2 的输入，图像数据维度大小和视频重建模型的滤波器尺寸也进行相应地调整。因为存在 3 个输入帧，模型新的过滤器尺寸为 $3\times f_1\times f_1\times C_1$，其中 C_n 表示第 n 层的内核数量，卷积层 L_2 的滤波器大小也扩大为 $3\times C_1\times f_2\times f_2\times C_2$，卷积层 L_3 的滤波器尺寸为 $3\times C_2\times f_3\times f_3\times 1$。

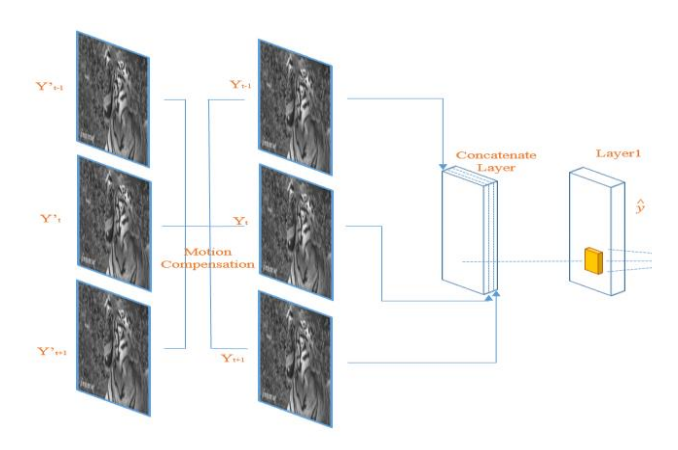


图 1　视频重建模型架构示意图

预训练模型和视频重建模型中的卷积层 L_1 的滤波器尺寸不同，视频重建模型中的第一维是预训练模型的三倍，因为三帧输入帧是沿着时间维度进行连接的。视频重建模型中使用的核宽度、核深度及核数量与预训练模型一致，以便预训练滤波器的值可以直接用于视频重建模型。唯一的区别是视频重建模型的卷积层 L_1 中的滤波器是使用三帧输入帧，所以在视频重建网络的连接层中的滤波器深度是在预训练模型中的三倍。

此外，卷积层 L_1 的输出数据类似于由单帧重建获得的输出数据，因为 L_2 和 L_3 保持与单帧重建模型中的相同设置。为了正确初始化视频重建模型，假设不是使用三个连续帧的视频，而是使用相同帧的三倍作为输入，视频重建模型和图像重建模型的输出结果应该是相同的。因为 L_2 和 L_3 在结构上和预训练重建架构中是相同的，所以只需确保两个模型的输入数据到卷积层 L_2 是相同的。对于预训练模型，由 H_1 表示 L_1 的输出数据具有尺寸 $M\times N\times C$，其元素 $h(i,j,c)$ 的计算为

$$h(i,j,c)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}w(m,n,t,c)y_t(i-m,j-n)+b(c) \tag{1}$$

其中：$w(m,n,t,c)$ 表示滤波器的权重；$b(c)$ 表示偏差；c 表示核的编号；y_t 表示时刻 t 下的输入帧，权重大小为 $M\times N\times1\times C$，第三维度尺寸为 1，因为在时刻 t 只有一帧输入图像，那么视频重建模型的相同数据计算为

$$h_v(i,j,c)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{t'=t-1}^{t+1}w_v(m,n,t',c)\times\hat{y}(i-m,j-n,t')+b_v(c) \tag{2}$$

其中：$w_v(m,n,t',c)$ 和 $b_v(c)$ 分别表示视频重建模型的权重和偏差；$\hat{y}$ 包含三个连续帧 $y(t-1)$、$y(t)$ 和 $y(t+1)$，将它们进行连接后可以根据输入图像 $y(t-1)$、$y(t)$ 和 $y(t+1)$ 表示为

$$\begin{aligned}h_v(i,j,c)=&\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}w_v(m,n,t-1,c)y_{t-1}(i-m,j-n)+\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}w_v(m,n,t,c)y_t(i-m,j-n)\\&+\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}w_v(m,n,t+1,c)y_{t+1}(i-m,j-n)+b_v(c)\end{aligned} \tag{3}$$

通过在式（3）中设置 $h_v=h$，并将 $y(t+1)$ 和 $y(t-1)$ 替换为 $y(t)$，若使式（1）与式（3）相等，需满足以下条件：

$$w(m,n,t,c)=w_v(m,n,t-1,c)+w_v(m,n,t,c)+w_v(m,n,t+1,c) \tag{4}$$

$$b(c)=b_v(c),\forall m,n,c \tag{5}$$

在实验中，本文将视频重建模型的滤波器权重和偏差初始化为

$$w_v(m,n,t-1,c)=w_v(m,n,t,c)=w_v(m,n,t+1,c)=\frac{1}{3}w(m,n,t,c) \tag{6}$$

$$b_v(c)=b(c),\forall m,n,c \tag{7}$$

这相当于在应用卷积层 L_1 之前对输入图像进行取平均操作。理想的运动补偿帧应该与其参考帧相同，使用这种框架训练神经网络理论上将导致 L_1 和 L_2 中的帧 $(t-1)$、帧 (t) 和帧 $(t+1)$ 为相等权重，特别是在视频帧内容中包含运动非刚性对象的情况下，会对重建效果造成较大影响。如果分别重建视频序列的每一帧，则每一帧将在某一时刻点处于当前时刻 (t)、之前时刻 $(t-1)$ 或后续时刻 $(t+1)$ 帧的位置。因此，从帧 $(t-1)$ 到当前帧 (t) 以及从帧 $(t+1)$ 到当前帧 (t) 的运动补偿误差应该是相同的。这意味着 L_1 中帧 $(t-1)$ 的滤波器和帧 $(t+1)$ 的滤波器权重应该是相同的。同理，L_2 中的所有过滤器权重也应该是相同的。因此，对模型中 L_1 的权重采用相同设置。在帧 $(t-1)$ 中的特定空间位置处应用相同的滤波器设置也可以表示相同的局部相关性，本质上在时间维度上扩展了网络的卷积性质，使滤波器共享相同的权重。

1.2　预训练模型

在开始对视频重建模型进行训练之前，首先对重建模型的权重进行了预训练。图像预训练模型仅具有卷积层，其优点是输入图像的尺寸可以是任何大小，并且算法不是基于图像块，

保证了算法的时间效率。其中，Y 表示输入的低分辨率图像，X 表示输出的高分辨率图像。预训练模型由三个卷积层组成，其中两个隐层 L_1 和 L_2 之后是线性修正单元（ReLU）。卷积层 L_1 由尺寸为 $1\times f_1\times f_1\times C_1$ 的滤波器组成，其中 $f_1\times f_1$ 表示核大小，C_1 表示 L_1 的核数量。卷积层 L_2 和 L_3 的滤波器尺寸分别为 $C_1\times f_2\times f_2\times C_2$ 和 $C_2\times f_3\times f_3\times 1$。$L_3$ 仅有一个核来获取图像作为输出；否则，需要具有一个内核的附加层，以及其他聚合等后处理步骤。

1.3　运动补偿

如果视频中发生较大的运动移位或运动模糊，运动补偿可能会很困难，这也可能导致高分辨率图像的重建过程中出现边界效应和伪像现象，从而削弱重建效果。模型引入自适应运动补偿方法，减少了在运动估计时相邻帧的影响，可根据式（8）进行运动补偿：

$$y_{t-T}^{amc}(i,j)=(1-r(i,j))y_t(i,j)+r(i,j)y_{t-T}^{mc}(i,j) \tag{8}$$

其中：$r(i,j)$ 控制每个像素位置 (i,j) 处的中心帧和相邻帧之间的相关性；$y(t)$ 是中心帧；y_{t-T}^{mc} 是运动补偿的相邻帧；y_{t-T}^{amc} 是应用自适应运动补偿后的视频帧；$r(i,j)$ 定义为

$$r(i,j)=\exp(-ke(i,j)) \tag{9}$$

其中：k 是常数参数；$e(i,j)$ 是运动补偿的误匹配误差。较大的误差可能是由于运动闭塞、物体模糊，或位置 (i,j) 接近运动边界而造成。根据式（8）和（9），当运动补偿误差在位置 (i,j) 处较大时，对应的权重较小。这意味着自适应运动补偿像素只是当前帧中的像素，从而更好地保证了重建帧的效果。

1.4　通道优化

一种提升图像分辨率的方法是在低分辨率空间中以 $\frac{1}{r}$ 的分数步幅进行卷积，随后在高分辨率空间中以步幅为 1 进行卷积。由于卷积操作发生在高分辨率空间中，算法计算开销将增加 r^2 倍。或者在具有尺寸为 k_s 的滤波器 W_s 的低分辨率空间中进行 $\frac{1}{r}$ 为步幅的卷积操作，激活卷积 W_s 的不同权重部分，恰好落在像素之间的权重不被激活和计算，激活模式的数量正好是 r^2。根据每个激活模式的位置，最多可以激活 $\lceil\frac{k_s}{r}\rceil^2$ 个权重。根据不同的子像素位置 $\mathrm{mod}(x,r)$、$\mathrm{mod}(y,r)$，滤波器在图像上的卷积操作周期性地激活这些模式，其中 x，y 分别表示高分辨率空间中的输出像素坐标。当 $\mathrm{mod}(k_s,r)=0$ 时，可采用如下方法实现：

$$I^{SR}=f^L(I^{LR})=PS(W_L*f^{L-1}(I^{LR})+b_L) \tag{10}$$

其中：$PS(*)$ 是将维度为 $H\times W\times C\cdot r^2$ 的张量元素重新排列成维度为 $rH\times rW\times C$ 的张量的周期性算子，在数学上可以描述为

$$PS(T)_{x,y,c}=T_{\lfloor x/r\rfloor,\lfloor y/r\rfloor,c\cdot r\cdot \mathrm{mod}(y,r)+c\cdot \mathrm{mod}(x,r)} \tag{11}$$

因此，卷积算子 W_L 具有形状 $n_{L-1}\times r^2C\times k_L\times k_L$，当 $k_L=\frac{k_s}{r}$ 和 $\mathrm{mod}(k_s,r)=0$ 时，它相当于具有滤波器 W_s 的低分辨率空间中的子像素卷积，输出层从低分辨率图像特征图直接生成一个用于每个特征图的放大滤波器，从而得到高分辨率图像。

对于目标函数，可由给定的高分辨率图像 $I_n^{HR},n=1\ldots N$ 组成的训练集，生成相应的低分辨率图像 $I_n^{LR},n=1\ldots N$，并计算重建图像的像素平均误差（MSE）作为训练网络的目标函数：

$$l(W_{1:L},b_{1:L})=\frac{1}{r^2HW}\sum_{x=1}^{rH}\sum_{x=1}^{rW}(I_{x,y}^{HR}-f_{x,y}^{L}(I^{LR}))^2 \tag{12}$$

2　实验结果与分析

首先将提出的模型与其他图像和视频重建算法进行了比较；然后定量地研究不同的视频重建架构的性能，以及预训练和运动补偿机制对视频重建效果的影响；最后，对实验结果进行比较与分析。

2.1　数据集

实验使用一个公开的视频数据库，该库包含 4 K 高分辨率（3840×2160 像素）的未经压缩的视频序列短片，选取视频中包含的 59 个场景帧，使用其中 53 帧进行训练和 6 帧进行测试，使用每个测试序列的 4 帧，并且计算 24 个测试帧的平均峰值信噪比（PSNR）值和结构相似性（SSIM）值作为性能指标。实验以 $\frac{1}{4}$ 为缩放因子对视频进行采样，得到 960×540 像素分辨率的图像，以便更好地与其他重建算法进行比较。实验还选择在另外一组视频集“Videoet4”上进行测试，实验过程中跳过每个视频的第一帧和最后一帧，以便使视频重建的过程中始终具备一整套 3 个连续帧作为模型输入。

2.2　模型参数设置

用于预训练的模型具有 3 个卷积层，其中 L_1 和 L_2 后面各自接 ReLU 单元，L_1 具有 64 个内核，内核大小为 9×9；L_2 具有 32 个内核，内核大小为 5×5；L_3 具有一个大小为 5×5 的内核。视频重建模型的滤波器具有与图像预训练模型相同的初始参数配置。

将图像和视频转换到 YCbCr 颜色空间，并且仅将亮度通道（Y 通道）用于训练、测试和客观性能指标计算。为了建立有效的视频训练集，从训练视频场景中提取了 3 组连续帧，使用 MATLAB 实现对所需要的缩放因子 2、3 或 4 进行抽样，并将所得到的低分辨率帧以双三次插值向上采样为原始分辨率。然后分别计算从第一帧和最后一帧到中心帧的光流，并计算得到运动补偿帧。从所得到的 3 帧（2 帧运动补偿帧和一帧中心帧）中提取 36×36×3 的数据立方块，即连续 3 帧的 36×36 像素图像块。如果帧中图像块不包含足够的结构信息，则将该帧进行移除，最终建立的训练数据库由图像数据立方体组成。

为了在训练阶段优化滤波器的权重和偏差，需要定义用于最小化的损失函数。训练数据集的输出图像和真值图像之间的欧氏距离需要测量，峰值信噪比（PSNR）的性能测量也与欧氏距离直接相关。为了避免训练期间的边界效应，可以对 36×36 像素的图像块采用零填充，或允许更小的卷积输出，即每个卷积层的输出图像块相应收缩。缩小的输出图像块对应于原始图像块的 20×20 中心像素，然后将这些中心像素用于计算损失函数。在视频重建的训练中，实验采用 240 批次，前两层的学

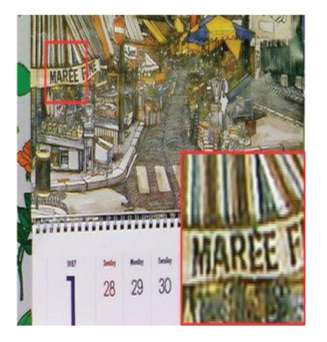
习率为 10^{-4}，最后一层的学习率为 10^{-5}，权重衰减率为 5×10^{-4}。使用预训练机制，进行 2×10^{5} 次迭代，不同帧的迭代滤波器效果如图 3 所示。实验表明，降低学习率对重建效果并没有进一步的改善，因此，实验在整个训练过程中保持学习率不变。



Original



Bicubic:23.819



Bayesian-MB:25.113



A+:25.291



SRCNN:25.526



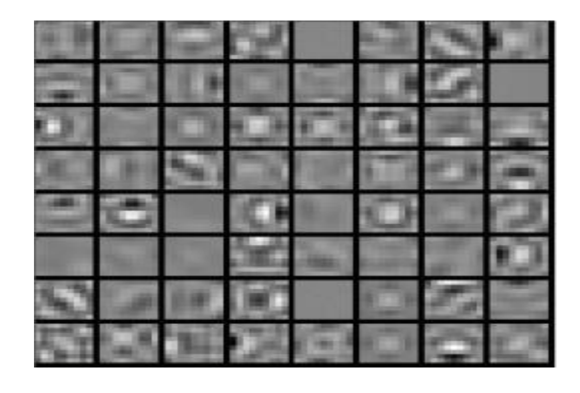
Ours:26.103

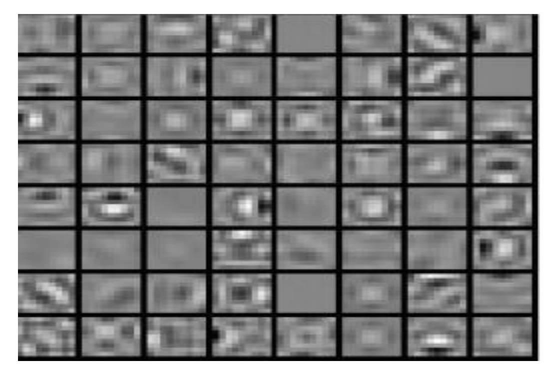
图 2 不同算法的性能比较（PSNR/dB）

2.3 与经典算法比较

将提出的模型分别与经典的单帧和视频重建算法进行比较，以双三次插值算法作为基准，可以通过简单地将单帧重建模型分别应用于每一帧来实现视频超分辨率重建，包括 Bicubic 插值重建算法、A+算法[9]以及 SRCNN 模型[17]。此外，还与经典的视频重建算法 Bayesian-MB[13]进行了比较，使用±1 帧相邻帧测试所有视频重建方法。

模型实现的第一个变化是使用 19×19 像素的双三次插值向上采样图像块，而不是使用 5×5 像素的输入图像块，这样可以更好地与使用双向插值输入的方法进行直接比较；第二个改变是用 ReLU[23]替代之前算法使用的 S 形激活函数，因为前者提供更快的模型收敛速度。





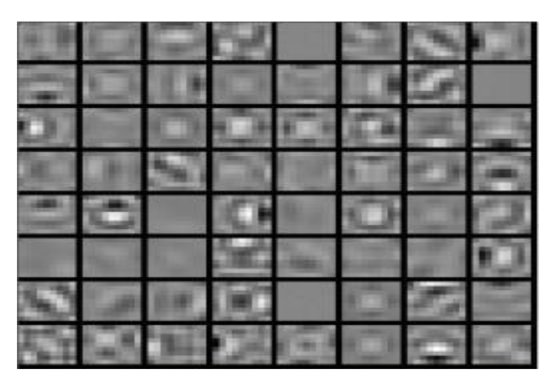


图 3 三帧训练滤波器的比较

表 1 不同重建方法的性能（PSNR）比较/dB

Dataset	scale	Bicubic	A+	SRCNN	Bayesian-MB	Ours
Videoset4	2	28.43	30.53	30.70	30.63	**31.2**
Videoset4	3	25.28	26.36	26.51	26.43	**26.66**
Videoset4	4	23.79	24.59	24.69	24.14	**24.7**

表 2 不同模型的客观评价指标 SSIM 比较

Dataset	scale	Bicubic	A+	SRCNN	Bayesian-MB	Ours
Videoset4	2	0.867	0.9154	0.917	0.923	**0.926**
Videoset4	3	0.733	0.7904	0.793	0.807	**0.807**
Videoset4	4	0.633	0.6889	0.692	0.687	**0.701**

图 4 所示为提出的模型在训练过程中的收敛速度与其他算法的比较。由图可以看出，与同样基于神经网络的 SRCNN 模型相比，该模型能够更快地收敛到稳定值；同时与其他所有重建算法相比，也能够更加高效地重建得到较高的 PSNR 值。除了模型训练效率高以外，在测试阶段，该算法是基于卷积神经

网络的前馈计算过程，大大减少了传统方法中由于大量迭代运算造成的重建效率损失，在硬件 GPU 加速下对于单帧视频图像的平均重建时间约为 0.26 s。在综合考虑重建精度和重建效率的情况下，对于不同的视频帧图像，该模型能够快速重建得到具有更高客观评价指标的重建图像，同时在主观视觉效果上能够有效去除重建图像的边缘和纹理模糊现象（图 5）。

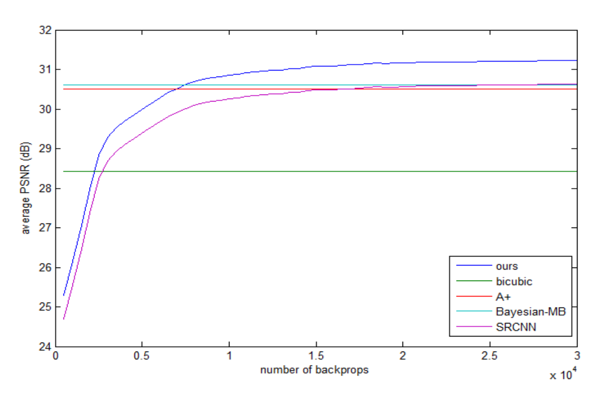


图 4 不同模型的重建效率收敛曲线比较

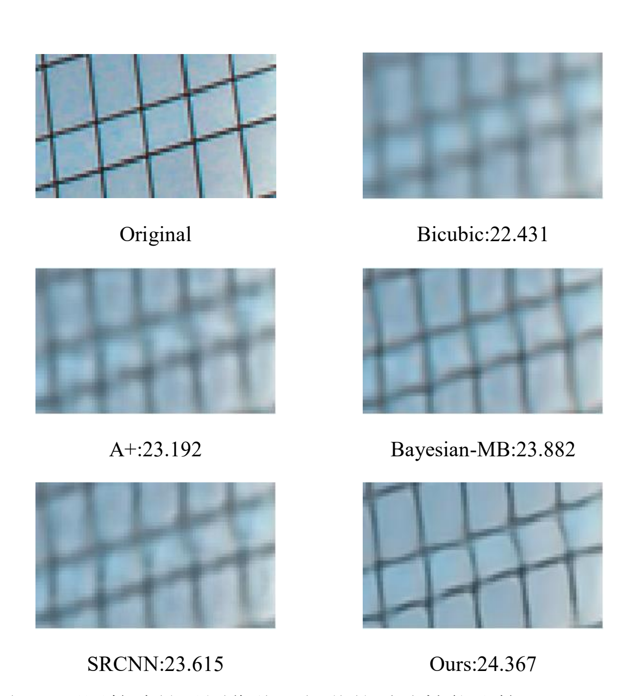


图 5 不同算法针对图像纹理细节的重建性能比较（PSNR/dB）

表 1 和 2 分别记录了不同的算法模型针对测试视频的客观评价指标 PSNR 和 SSIM 值。由表可以看出，所提出的模型在所有实验中均得到了较高的平均指标。在测试数据集上，与 SRCNN 模型相比，对于放大因子 2、3 和 4 均有提升，特别是对于高分辨率的重建，在 Videoet4 数据集上对放大因子为 2 和 3 的增益比放大因子为 4 时更加明显。实验结果表明，视频重建质量的提升对于低放大因子越明显，这可能是由于模型中更精确的运动补偿算法。图 2 和 5 展示了使用不同算法重建视频测试集的示例及其对应的 PSNR 值。与其他算法相比，所提出的重建模型能够较好地恢复更多的图像细节内容，明显去除了图像边缘的模糊效应，进一步提升了图像的细节清晰度，为后续相关的图像处理任务提供了更好的支持。

3 结束语

本文充分利用视频多帧图像间的空间和时间信息，提出了一种新的基于卷积神经网络的视频重建模型。通过比较不同的模型结构，对模型的输入帧进行运动补偿处理和预训练策略，能够有效提高视频重建的质量并减少训练时间。针对视频中的快速移动对象，引入自适应运动补偿方案来处理由此引发的运动模糊现象。实验表明，与经典方法相比，所提出的模型能够在视频图像重建中获得更优的客观评价指标和更高的图像质量，为其他需要高分辨率图像的视觉任务和应用场景提供了基础。对模型的进一步改进和优化，特别是针对特定应用场景的视频数据进行超分辨率重建及视觉应用，也是今后的研究方向之一。

参考文献：

[1] 何小海，吴媛媛，陈为龙，等．视频超分辨率重建技术综述 [J]．信息与电子工程, 2011, 9 (1): 1-6.

[2] Belekos S P, Galatsanos N P, Katsaggelos A K. Maximum a posteriori video super-resolution using a new multichannel image prior [J]. IEEE Trans on Image Process, 2010, 19 (6): 1451-1464.

[3] Babacan S D, Molina R, Katsaggelos A K. Variational Bayesian super resolution [J]. IEEE Trans on Image Process, 2011, 20 (4): 984-999.

[4] Liu Ce, Sun Deqing. On Bayesian adaptive video super resolution [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2014, 36 (2): 346-360.

[5] Yang J, Wright J, Huang T, et al. Image super-resolution via sparse representation [J]. IEEE Trans on Image Process, 2010, 19 (11): 2861-2873.

[6] Aharon M, Elad M, Bruckstein A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation [J]. IEEE Trans on Signal Process, 2006, 54 (11): 4311-4322.

[7] Song B C, Jeong S C, Choi Y. Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the-fly dictionary training [J]. IEEE Trans on Circuits and Systems for Video Technology, 2011, 21 (3): 274-285.

[8] 张岩，李建增，李德良，等．无人机侦察视频超分辨率重建方法 [J]．中国图象图形学报, 2016, 21 (7): 967-976.

[9] Timofte R, De Smet V, Van Gool L. A+: adjusted anchored neighborhood regression for fast super-resolution [C]// Proc of the 12th IEEE Asian Conference on Computer Vision. 2014: 1920-1927.

[10] Schulter S, Leistner C, Bischof H. Fast and accurate image upscaling with super-resolution forests [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3791-3799.

[11] Glasner D, Bagon S, Irani M. Super-resolution from a single image [C]// Proc of IEEE International Conference on Computer Vision. 2009: 349-356.

[12] 覃凤清，何小海，陈为龙，等．一种基于子像素配准视频超分辨率重建方法 [J]．光电子．激光, 2009, 20 (7): 972-976.

[13] Ma Z, Jia J, Wu E. Handling motion blur in multi-frame super-resolution [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition.

2015: 5224-5232.

[14] Takeda H, Milanfar P, Protter M, et al. Super-resolution without explicit subpixel motion estimation [J]. IEEE Trans on Image Process, 2009, 18 (9): 1958-1975.

[15] Shi W, Caballero J, Huszar F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1874-1883.

[16] Kappeler A, Yoo S, Dai Q, et al. Video super-resolution with convolutional neural networks [J]. IEEE Trans on Computational Imaging, 2016, 2: 109-122.

[17] Dong Chao, Loy C C, He Kaiming, et al. Learning a deep convolutional network for image super-resolution [C]// Proc of IEEE European Conference on Computer Vision. 2014: 184-199.

[18] Wang Zhangyang, Yang Yingzhen, Wang Zhaowen, et al. Self-tuned deep super resolution [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-8.

[19] Cui Z, Chang H, Shan S, et al. Deep network cascade for image super-resolution [C]// Proc of IEEE European Conference on Computer Vision. 2014: 1-16.

[20] Cheng Minghui, Lin Naiwei, Hwang K, et al. Fast video super-resolution using artificial neural networks [C]// Proc of the 8th International Symposium on Communication Systems, Networks & Digital Signal Processing. 2012: 1-4.

[21] Liao R, Tao X, Li R, et al. Video super-resolution via deep draft-ensemble learning [C]// Proc of IEEE IEEE International Conference on Computer Vision. 2015: 531-539.

[22] 黄璇，杨晓梅. 基于低秩及全变分的视频超分辨率重建 [J]. 计算机应用研究, 2015, 32 (3): 938-941.

[23] Maas A, Hannun A, Ng A. Rectifier nonlinearities improve neural network acoustic models [C]// Proc of IEEE International Conference on Machine Learning. 2013: 1-8.